Base de données et vocabulaires contrôlés

Utilisation des vocabulaires contrôlés en ligne pour les données

JDEV 2017 (05/07/2017)

Marie-Claude Quidoz – Baptiste Laporte

Contexte

- Thèse de MA Laporte (2011) en écoinformatique

- Groupe interopérabilité de rBDD / CESAB (2013-2016)
  - Informaticien / documentaliste /chercheur
  - http://rbdd.cnrs.fr/spip.php?rubrique34

- GDR Semandiv (2017-2021)
  - Informaticien / documentaliste /chercheur
  - Axe 4 : Visualisation, interrogation, mise en correspondance avec les bases de données

Pourquoi ?

- Contrôle des données
- Interopérabilité
- Evolution des données

Contrôle des données

Réduire les erreurs humaines

Réduire l'hétérogénétité syntaxique

Réduire l'hétérogénéité sémantique

BDD propre = basé sur une sémantique définie !!!

Nécessité de l'interopérabilité

- mise en commun de données issues d'observations/expérimentations conçues initialement de façon indépendante et/ou dans des contextes disciplinaires différents
- Produire de nouvelles connaissances à partir de ces données
- Interface entre écologie et sciences de l'information: développement de l'écoinformatique

Interopérabilité

Association données/sémantique commune

=

premier pas

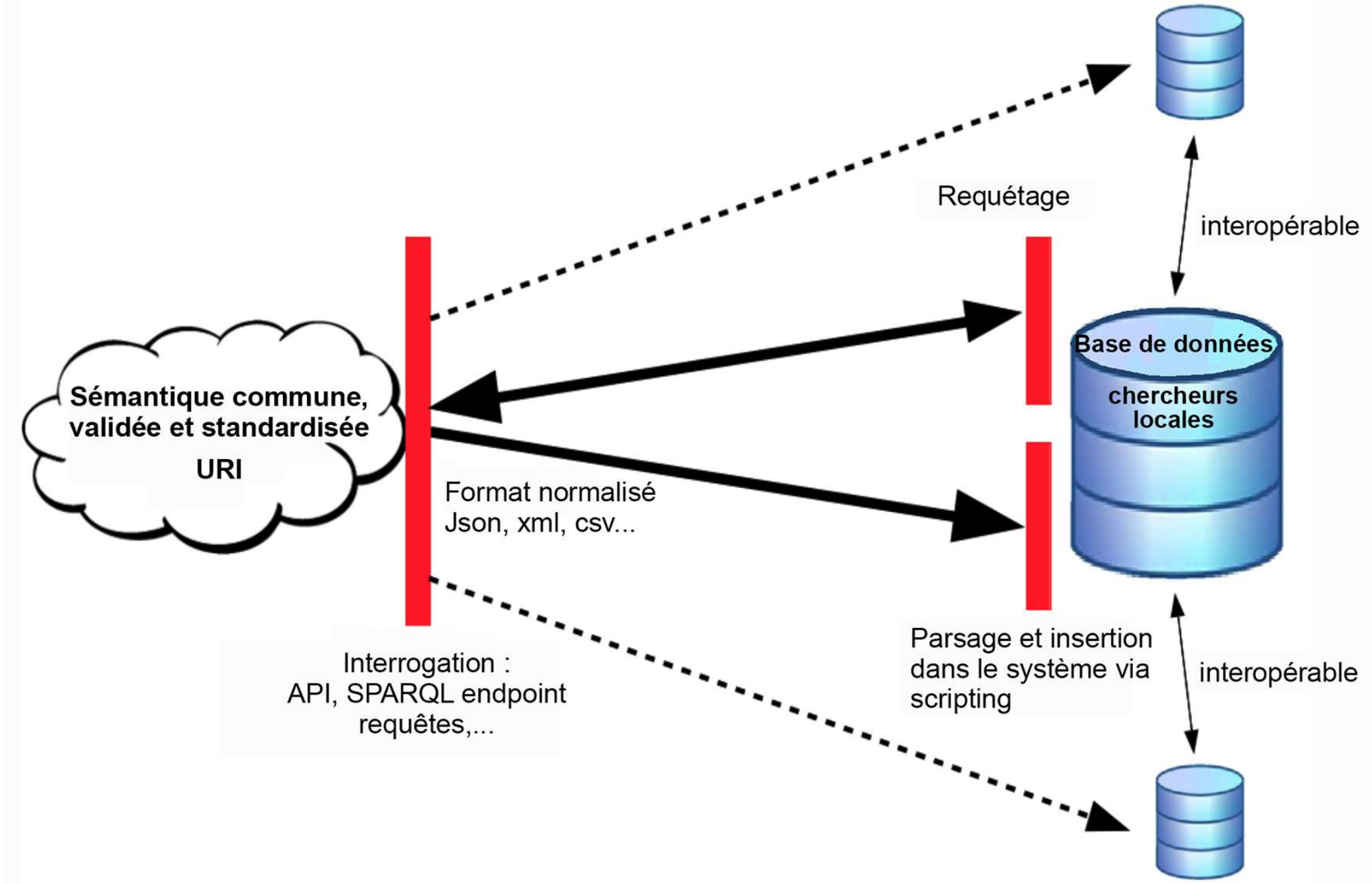
vers l'interopérabilité

Evolution des données

Web sémantique = données intelligentes

URI/URL (persistant) + lien entre eux = évolution

Synonymie, hiérarchie, etc.


Requétage
interopérable
Sémantique commune,
validée et standardisée
URI
Base de données
chercheurs
locales
Format normalisé
Json, xml, csv...
Interrogation :
API, SPARQL endpoint
requêtes,...
Parsage et insertion
dans le système via
scripting
interopérable

Exemple d'utilisation

Utilisation d'un thésaurus

Simplicité de construction, outil thesauform existant

Lien avec une base de données SQL (EAV)

SQL = contrôle fort des données

EAV = flexibilité de la base, permet l'annotation

Interfaçage avec un langage de scripting PHP

Librairie lecture json disponible, etc

Modèle EAV

Schéma sensible à modifications

Ne pas coder les attributs des entités en dur

coder le couple nom de l'attribut/valeur

⇒ Ajouts de nouveaux attributs sans modifier le schéma conceptuel / la structure des requêtes pour extraire l'information

Inconvénients :

- impossible de mettre des contraintes physiques dessus sans vérification poussés des données
- le type des données non spécifié = colonne supplémentaire
- Ralentissement des performances
- la lecture de la donnée en ligne = transformation des résultats des requêtes
- Requêtage direct des données difficile, besoin d'interface (pl-SQL, scripting externe)

Thesaurus

- *Une liste de terme définie permettant d'être sûr de travailler sur la même chose entre expert et non-expert, début de l'interopérabilité.* (**définition personnelle sûrement erronée**)
- *Liste organisée de termes contrôlés et normalisés représentant les concepts d'un domaine de la connaissance*

Les termes sont reliés entre eux par des relations de :

- Synonymie (terme équivalent)
- Hiérarchie (terme générique et spécifique)
- Association (terme associé)

Capacité d'indexation et de recherche

Thesaurus

Pour une expérience réussie

- couverture du domaine
- expertise des validateurs
- respect des standards
- utilisation par une communauté

Les outils

- Différentes solutions
  - licence (libre/commercial)
  - utilisateur impliqués (mono poste / collaboratif)
  - différents standards (OBO, OWL, SKOS, …)
  - différentes technologies (java, technologies web, …)
  - différents moyens d'accès (local, web, web sémantique compliant, etc.)

⇒ Intégration donnés = scientifiques utilisateurs (ergonomie importante)

Le Thesauform

Prototype développé dans le cadre d'une thèse en écoinformatique (2011)

- Collaboratif (web)
- Standard SKOS
- Simple d'utilisation
- Visualisation travaillée
- Extraction par API

=> https://github.com/CESAB-FRB/Thesauform

Technique

J2EE, modèle MVC

Fichier texte OWL compatible SKOS

JENA

Exemples d'utilisations réussies

A Terminological Resource for Plant Functional Diversity : http://top-thesaurus.org/

Thesaurus for Soil Invertebrate Trait-based Approaches : http://tsita.cesab.org/

Démonstration

Thesauform

http://localhost:8080/thesauform_V1.0

Thesauform API

http://t-sita.cesab.org

Exemple d'API:

Id = 1 : test existence of trait
Id = 2 : select all direct sons of a trait
Id = 3 : all qualitative traits
Id = 4 : all quantitive traits
Id = 5 : select trait real name
Id = 6 : test a trait is quantitative
Id = 7 : get synonym of a trait
Id = 8 : get all trait with unit interval (temporal traits)
Id = 9 : get all trait and synonyms with at least one synonym
Id = 10 : test a trait is interval

http://t-sita.cesab.org/BETSI_php.jsp?id= (3,4,8,9)
http://t-sita.cesab.org/BETSI_php.jsp?id= (1,2,5,6,7,10)&trait=Repulsive_glands

Scripting PHP

Curl et json: apiSPARQL.php

Gestion des exceptions: insertTraits.php

Requête: exempleSyn.php

Limitations

Besoin de connexion

Risque de création de multiples thésaurus

Gestion de l'ajout des termes utilisateurs spécifiques